Epigenetic Tracking: a Model for Multicellular Biology

Alessandro Fontana

email address: afalex026@gmail.com

April 9, 2013

Abstract

Epigenetic Tracking is a model of systems of biological cells, able to generate arbitrary 3-dimensional cellular structures of any kind and complexity (in terms of shape, number of cells, etc.) starting from a single cell. If we interpret such structures as a metaphor for biological organisms, we can conclude that this model has the potential to reproduce the complexity typical of living beings. It can be shown how the model is able to mimic a simplified version of key biological phenomena such as development, the presence of junk DNA, the phenomenon of ageing and the process of carcinogenesis. The model links properties and behaviour of genes and cells to properties and behaviour of the organism, describing and interpreting the said phenomena with a unified framework: for this reason, we think it can be proposed as a model for multicellular biology. The material contained in this paper is not new: the model and its implications have been described previously. The objective of this work is to present the different aspects of the theory with a unified approach. The paper is divided into six parts: the first part is the introduction; the second part describes the cellular model; the third part is dedicated to the evo-devo process and transposable elements; the fourth part deals with junk DNA and ageing; the fifth part explores the topic of cancer; the sixth part draws the conclusions.

Contents

1 Introduction

1.1 Embryogenesis and Artificial Embryology

Embryogenesis, the process by which the zygote develops into a progressively more complex embryo to become an adult organism, is one of the greatest miracles of nature, yet poorly understood. Such process is known to be guided by the DNA contained in the zygote, which is copied unaltered at each cell division. How does the DNA organise the growth process? Since all cells have exactly the same genetic makeup, how can each cell know which type of specialised cell it is destined to become? How are 3-dimensional structures generated from a linear DNA code? How do cells coordinate themselves to behave like an organism? These are some of the questions waiting for an answer.

Artificial Embryology is a sub-discipline of Artificial Life aimed at modelling the process of morphogenesis and cellular differentiation that drives embryogenesis. Models in the field of Artificial Embryology [18] can be divided into two broad categories: grammatical models and cell chemistry models. In the grammatical approach development is guided by sets of grammatical rewrite rules; a prototypical example of grammatical models is represented by L-systems, first introduced by Lindenmayer [19] to describe the complex fractal patterns observed in the structure of trees. The cell chemistry approach draws inspiration from the early work of Turing [29], who introduced reaction and diffusion equations to explain the striped patterns observed in nature; this approach attempts to simulating cell biology at a deeper level, reconstructing the networks of chemical signals exchanged within and between cells.

1.2 Approach followed in constructing the model

This work is concerned with an Artificial Embryology model called *Epigenetic Tracking (ET)*, described in [4], [5], [6], [7], [8], [9], [10], able to generate arbitrary 3-dimensional cellular structures starting from a single cell. The model has been constructed with the following approach: first, we have designed the architecture of the model at a high level, based on known biological elements. Subsequently, we have added additional elements, not necessarily known, in order to "make things work in silico" (i.e. to produce interesting results in computer simulations). Finally, we have come back to biology, trying to guess which biological molecules play the role of the additional elements. As a consequence the model contains ingredients not necessarily adherent to current knowledge, but which can become a suggestion for biologists to look into new, previously unexplored directions.

Computer simulations have proved the capacity of the model to generate structures of any kind and complexity (in terms of shape, number of cells) starting from a single cell. If we interpret such structures as a metaphor for biological organisms, we can conclude that this model has the potential to reproduce the complexity typical of living beings. Furthermore, it can be shown how the model is able to reproduce a simplified version of key biological phenomena such as development, the presence of junk DNA, the phenomenon of ageing and the process of carcinogenesis. To our knowledge, this is the only model able to describe and interpret the said phenomena with a unified framework: for this reason, we think it can be proposed as a model for multicellular biology.

The material contained in this paper is not new: the model and its implications have been described previously. The objective of this work is to present the different aspects of the theory with a unified approach. The paper is divided into six parts: this first part is the introduction; the second part describes the cellular model; the third part is dedicated to the evo-devo method and transposable-elements; the fourth part deals with junk DNA and ageing; the fifth part explores the topic of cancer; the sixth part draws the conclusions and outlines future directions.

2 The cellular model of development

2.1 Driver cells, the genome and the epigenome

In this model phenotypes are represented as *cellular structures*, aggregates of cube-shaped cells deployed on a grid. Cells belong to two distinct categories: *normal cells*, which make up the bulk of the structure and *driver cells*, which are much fewer in number (by orders of magnitude) and are evenly distributed in the structure volume. If we compare normal cells to simple soldiers of an army, driver cells can be compared to their sergeants: in other words, driver cells have the power to issue orders that normal cells have to obey. The outcome of these orders is the orchestration of developmental acts called *change events*, which consist in the creation and deletion of large numbers of cells.

The former statement should not interpreted too rigidly: we are aware of the complexity and plasticity of biological systems. The message we intend to convey is that the *key orders* to produce change events are *initiated* by driver cells. Normal cells should not be considered mere passive responders and the presence of modulatory signals sent by normal cells "in the opposite direction" is by no means in contrast with this model. Moreover, as we shall see, cellular roles are not fixed: their assigment is the outcome of a dynamic process.

Development (Fig. 1) starts with a single cell (called zygote) placed in the middle of a grid and unfolds in N *developmental stages*, counted by a *global clock (GC)* shared by all cells. During development, a structure can be "viewed" in two ways: in *external view* colours represent cell types; in *internal view* colours represent cell states: blue is used for normal cells alive, orange for normal cells just created (i.e. in the current stage), grey for cells that have just died, yellow for driver cells (regardless of when they have been created).

The genetic information (Fig. 2), encoded as in nature by sequences of four numbers (0,1,2,3), is composed of two sets:

- the *fixed genome* and
- the *variable genome* or *epigenome*.

The fixed genome, as the name implies, is equal in all cells; the variable genome, on the other hand, can be different in different cells, reflecting the differentiation occurring during development.

The fixed genome is structured as an array of genes: each gene is composed of a left (or "if") part, encoding a condition, and a right (or "then") part, encoding an action: once the condition is verified, the action is executed. There are two types of genes:

- *developmental genes*, responsible for the changes occurring to the organism, and
- *metabolic genes*, responsible for cellular behaviour at steady state.

The variable genome, as we said, is composed of the elements which can become different in different cells. These elements can be grouped in four categories:

- the *driver state*, indicating whether a cell is driver or not;
- the *mobile code (MOC)*, an abstraction for all regulatory elements (proteins, RNA's) present in the cell;
- the *epigenetic marks* on metabolic genes, indicating whether they are structurally inactive or not;
- the *activation marks* on developmental genes, indicating whether they have been activated in the cell's lineage.

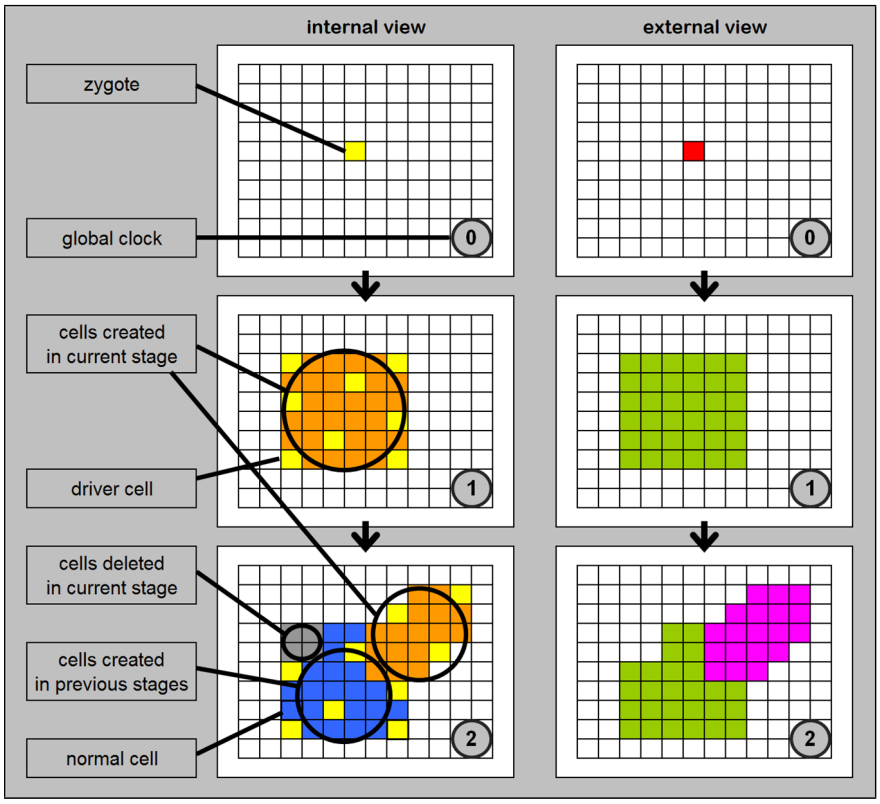

Figure 1: Development and views. Development starts with a single cell and unfolds in N developmental stages, counted by a global clock. During development, a structure can be seen from an internal and from an external perspective.

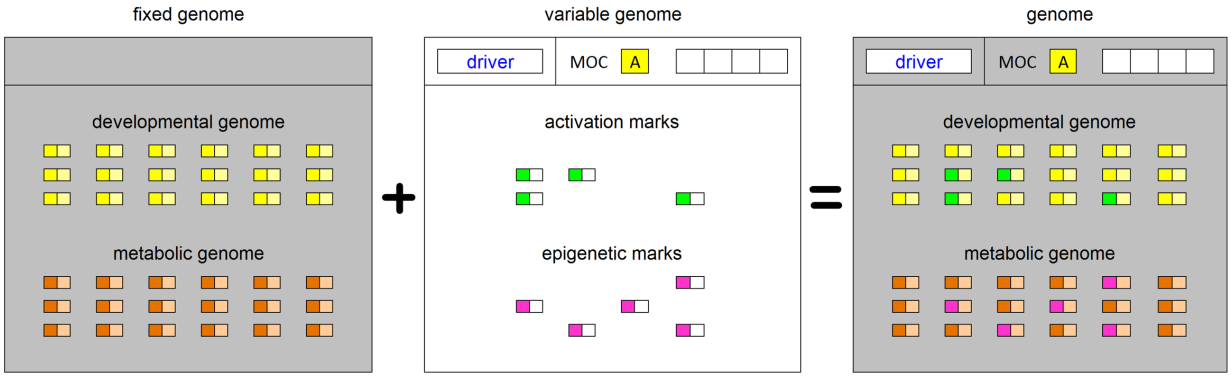

Figure 2: Fixed and variable genome. The fixed genome is equal in all cells, while the variable genome becomes different in different cells, reflecting the differentiation occurring during development. The fixed genome is composed of developmental genes (marked in yellow) and metabolic genes (marked in brown). In the variabel genome, the epigenetic marks render metabolic genes active or inactive; the activation marks indicate all developmental genes activated during development in the cell's lineage; the MOC is an abstraction for all regulatory factors (RNA, proteins) present in the cell; the driver mark indicates whether the cell is driver or not.

Developmental genes can be compared to "macro" biological genes or sets of genes co-regulated. The left part of a developmental gene is composed of the following elements: i) a field called *switch (SWC)*, indicating whether the gene can potentially be activated or not; ii) a field called *mobile sequence (MOS)*, than can match with the MOC; iii) a field called *timer (TM)*, than can match with the clock. At each developmental stage (marked by a different clock value), for each developmental gene and for each driver cell, it is checked if the gene's MOS matches the driver's MOC and if the gene's timer matches the clock (MOC and MOS sequences are indicated with letters: if a MOC and a MOS are labelled with the same letter, it means they match). If both conditions are verified, the right part of the gene is executed.

The right part of a developmental gene encodes a change event (Fig. 3). Two types of events are foreseen: *proliferation events* cause the activated driver cell (called *mother driver cell* or simply *mother cell*) to proliferate in the volume around it (called *change volume*); *apoptosis events* cause cells in the change volume to be deleted from the grid. The right part specifies also the shape of the change volume (and its colour in case of proliferation), in which the event takes place. Finally, a field of the right part "influences" the switch of some metabolic genes, turning them on or off and allowing differentiation to take place in the cell's metabolic network.

We can imagine that a proliferation event starts by producing a pool of undifferentiated, stem-like cells, based on the footprint of the driver cell activated: these cells are then induced to differentiate. This process could be implemented in nature through a class of molecules called **growth factors**, which during embryonic development act locally, either as paracrine or autocrine regulatory chemical messengers, as important regulators of cellular proliferation and differentiation [26].

Metabolic genes have a structure which closely resemble that of real genes. Taken together, metabolic genes compose a gene regulatory network which, thanks to the selective activation / deactivation of individual genes, becomes different in different cells, allowing specialisation to take place in the network. The selective activation / deactivation of metabolic genes is done changing the value of their epigenetic mark, an action carried out by developmental genes. Since the focus of this paper is on developmental events, metabolic genes will be omitted from this point onwards.

Key elements of this model are the clock and the MOC, which allow the activation of different portions of the developmental genome in a spatio-temporal specific manner. These elements, therefore, represent the main source of differentiation during development. The clock and the MOC can be interpreted in biology (Fig. 4) as the set of all regulatory factors present in the cell: for simplicity, we can think of them as a set of transcription factors (proteins). The timer and the MOS correspond to the gene regulatory sequences to which the factors bind. In our implementation the clock and the timer are represented as numbers, while the MOC and the MOS are represented as sequences of numbers (each number in the sequence can be interpreted as a transcription factor or a regulatory locus).

2.2 Generation of new driver cells

During a proliferation event, initially only normal cells are created in the change volume. On the other hand, the presence of a uniform distribution of driver cells in the structure volume throughout development is a key feature of this model. Indeed, if a part of the structure remained deprived of driver cells, no change event could occur in that part and further development would be impeded. Therefore new driver cells need to be created in the change volume. A further requirement is that each driver cell must have a distinct MOC value, to allow for differentiation.

In nature, a key role in cellular differentiation is played by a class of chemical compounds

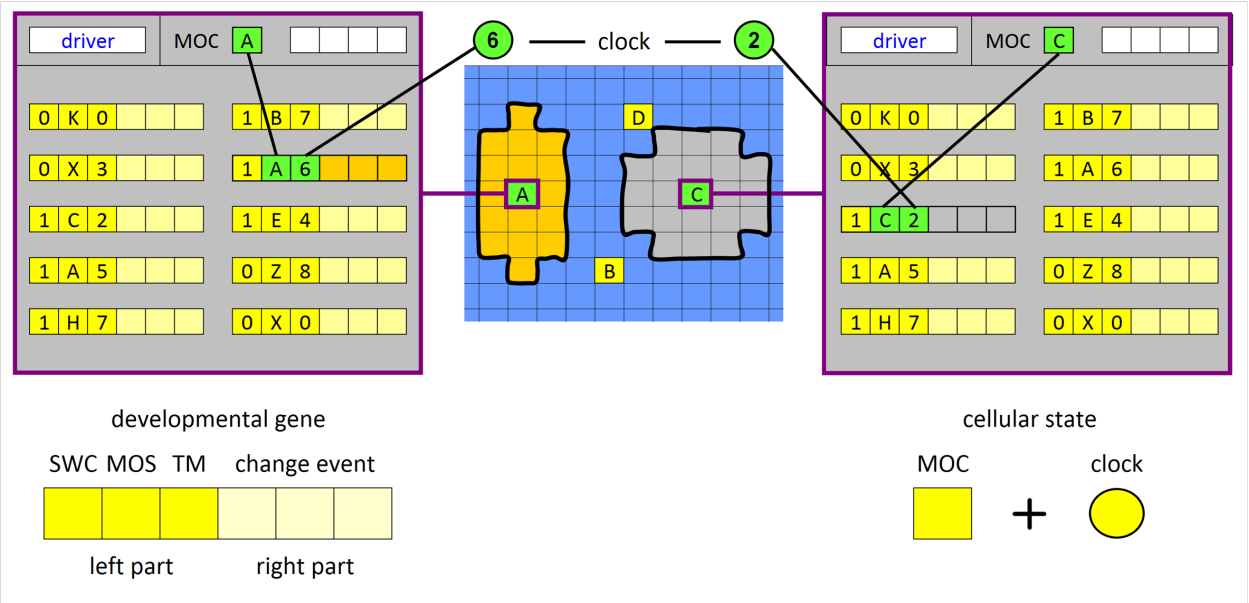

Figure 3: Change events. On the left a change event of type proliferation. MOC value A matches with MOS sequence A and the clock value matches with the timer value (6): since the switch is on (1), the gene is activated. The right part codes for an ellissoid-shaped proliferation event. Before proliferation, the MAD field of the gene changes the switches of some metabolic genes, turning them on or off. Such pattern of metabolic gene activation is inherited by all newly generated cells. On the right a change event of type apoptosis. For simplicity only the development genome is shown.

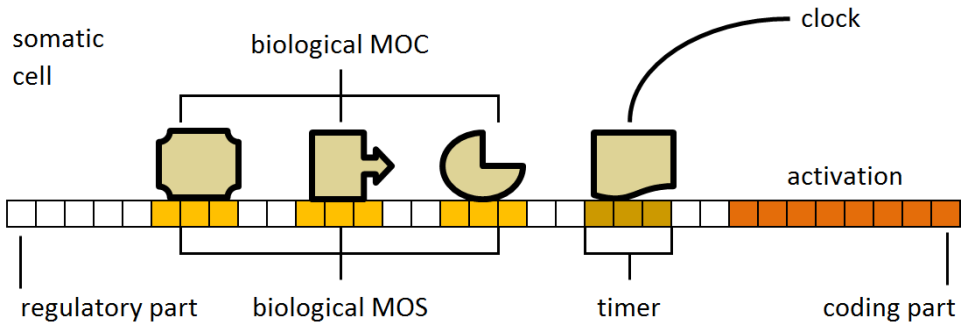

Figure 4: Biological MOC, MOS, clock and timer. The figure shows the components of a biological MOC and a biological clock (transcription factors) binding to the components of a biological MOS and a biological timer (regulatory sequences of a gene). As a result, the gene is activated (in general, a biological MOC can comprise transcription factors relevant to several genes).

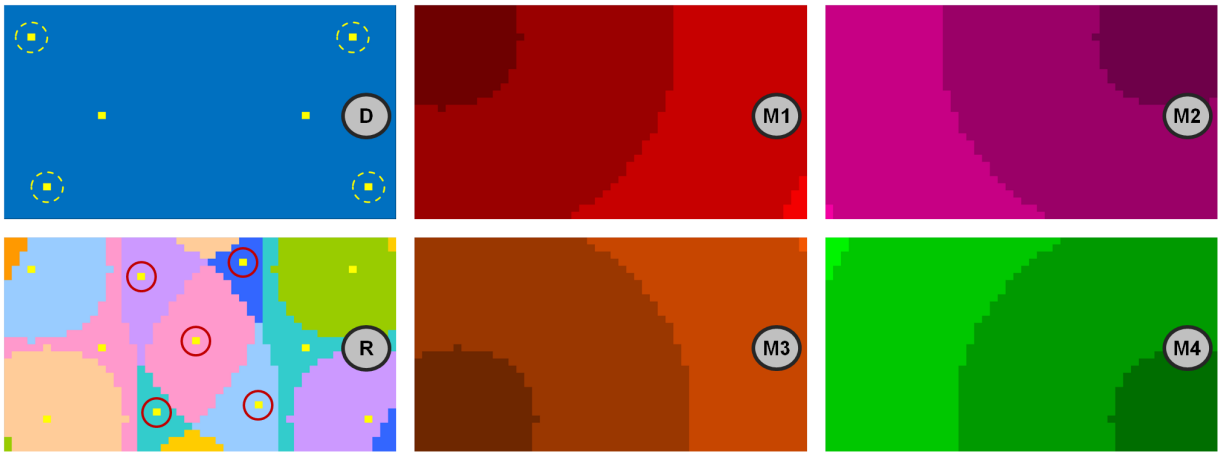

Figure 5: Simulated morphogens. Panel D shows a hypothetical distribution of driver cells. Some drivers emit signals of a finite number of types (4 in this example). Panels M1-M4 show the concentrations of morphogens of type M_1 (orange), M_2 (green) M_3 (purple) and M_4 (red). In panel R the different combinations of the concentration values of the signals are shown to partition the space into regions (marked with different colours). In each region new driver cells are created (in circles).

called **morphogens**. Morphogens are substances involved in the patterning of tissue development and the positioning of the various specialised cell types within a tissue. They are spread from a localised source and provide spatial information by forming a concentration gradient and inducing the expression of specific genes at distinct concentration thresholds. Different combinations of morphogens induce the emergence of new cell types through a characteristic "morphogenetic code" [15]. Well-known morphogens include Notch, Wnt, Hedgehog, and TGF-beta. Not surprisingly, genetic pathways exist dedicated to the processing of each of these morphogens. Each pathway starts when a chemical messenger binds to a receptor at the cell surface, and ends when downstream molecules enter the nucleus and participate in the regulation of target gene expression.

In ET the process of cellular differentiation is achieved through the presence of simulated morphogens, acting with the following mechanism. Some driver cells (Fig. 5) become sources of specific morphogens (of k distinct types: M_1, ... M_k). These signals partition the space into regions, each characterised by a distinct combination of signal types and strenghts. For example, all cells in a given region "sense" morphogen of type M_1 with strength 7, morphogen of type M_2 with strength 2, etc. In each region a normal cell is selected and turned into a driver.

To each new driver cell a new and unique MOC value must be assigned. When the driver is created, it has the same MOC value as the normal cell from which it was obtained, which coincides with the MOC value of the normal cell's mother. The assignment of a new MOC value is achieved through the processing of the signals received by a device called *MOC generator* (Fig. 6). The result is a new number which is appended at the end of the existing MOC value to produce the new MOC value for the driver cell. In biological terms, this corresponds to creating a new transcription factor, which is added to the repertoire of factors already present in the cell. Figure 7 shows an example of development for a hypothetical "organ", which alternates the occurrence of change events and the creation of new driver cells based on the mechanism described.

We hypothesise that the MOC generator of Fig. 6, which supervises the creation of the new element for the MOC value, is implemented in nature by all pathways dedicated to the

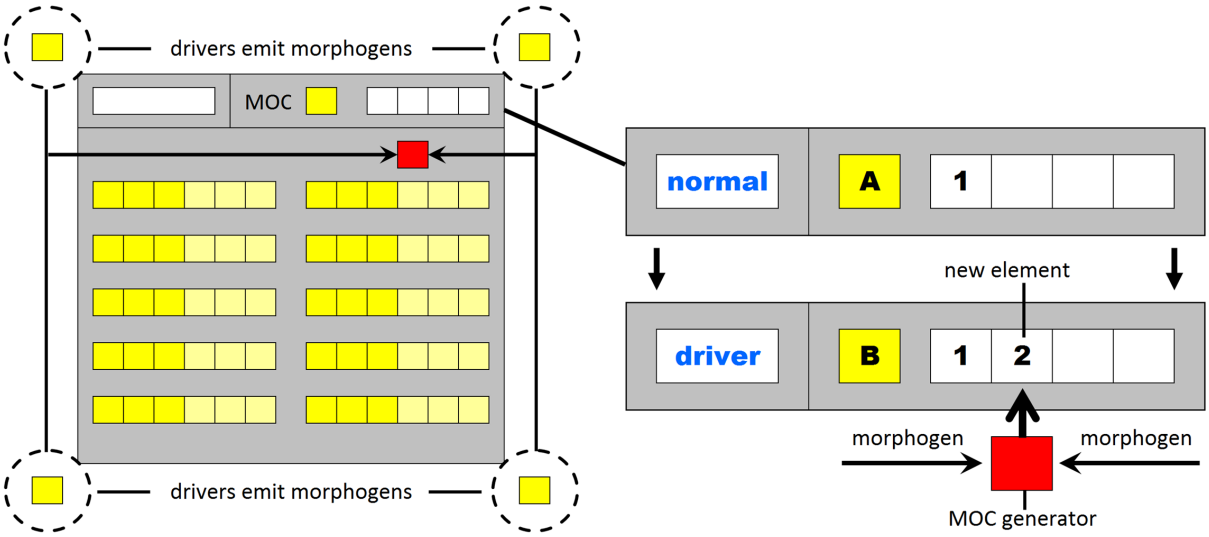

Figure 6: Generation of a new MOC value in a newly formed driver cell. Morphogens, released by some driver cells, reach a device called MOC generator, which outputs a new number (corresponding to a new transcription factor), added to the repertoire of numbers already present in the cell. This device is hypothesised to be composed of the genetic pathways dedicated to the processing of morphogens.

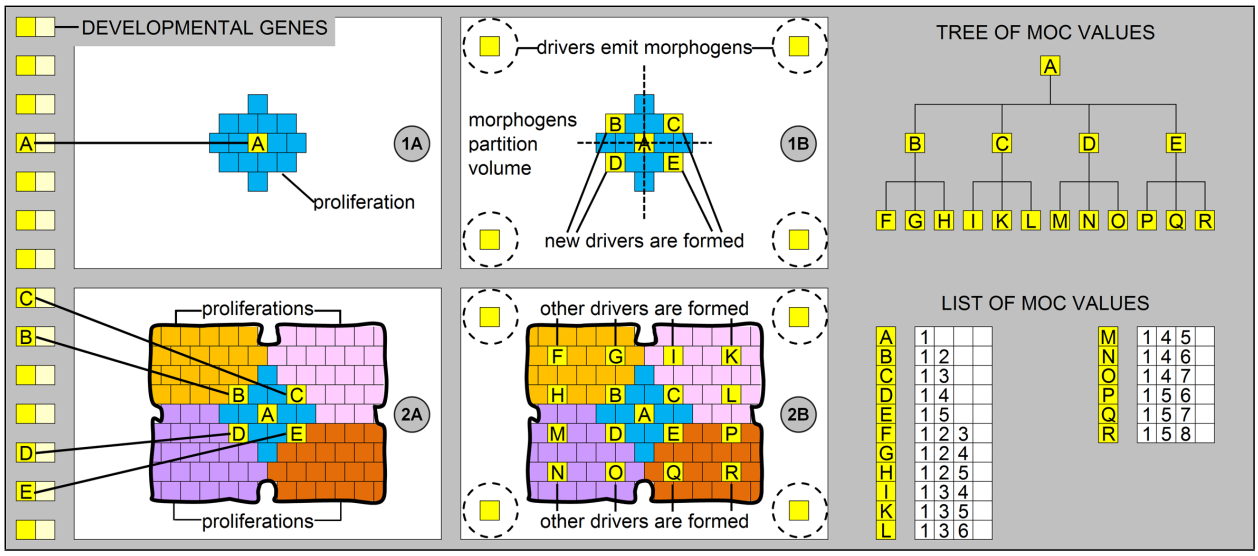

Figure 7: Creation of new drivers during development. The Figure shows the development of a hypothetical "organ" in two stages, carried out by five developmental genes (driver cells are marked in yellow, normal cells are shown with different colours, each corresponding to a particular differentiation state). The right part of the Figure reports the MOC values generated during development, which are organised in a tree-like structure. Each stage is shown in two panels, A and B. Panel A shows the occurrence of one or more change events, panel B shows the creation of new driver cells. This process is based on morphogens emitted by some driver cells (indicated with dashed circles in the corners), which partition the space in regions. One normal cell in each region is turned into a driver, and obtains a distinct MOC value. Blue normal cells, generated by driver cell A, are immature cells with a low degree of differentiation, while orange, pink, purple and red cells, generated by drivers B, C, D and E respectively, are mature, fully-differentiated cells.

processing of biological morphogens. The combined action of the pathways would result in the formation in the nucleus of the new driver cell of a new transcription factor. This could be done in the following way. Each pathway "senses" the strength of the relevant morphogen signal, converts it into a molecular signal, which is transduced to the nucleus. Here the interplay of such signals results in the creation of a new transcription factor (in biological cells the process could produce multiple new transcription factors). For example, if the combination of morphogen strengths is [2130] (i.e. Notch is received with strength 2, Wnt with strength 1, Hedgehog with strength 3 and TGF-beta with strength 0), transcription factor X is produced. If the combinations of the same signal strenghts is [3110], transcription factor Y is produced. These transcription factors are added to the repertoire of factors already present in the cell, and concur to determine the cellular state.

The distinction between normal cells and driver cells is the most important feature of this model. In section 2.1 we have shown how driver cells have the power the proliferate and generate normal cells. In this section we have described how new driver cells can be generated from normal cells, implying that the normal-driver transition is a two-way street. These two phenomena can be thought as two sides of a fundamental biological principle, that induces cellular systems to self-organise in a hierarchical fashion. We think of this *biological hierarchical principle* as a pervasive mechanism underlying all manifestations of cellular life, both in physiological and pathological conditions.

2.3 Biological interpretation of driver cells

In biological development a pivotal role is known to be played by **stem cells**, a class of cells found in most multi-cellular organisms. *Embryonic stem cells* (found in the inner cell mass of the blastocyst) are totipotent cells, which means that they are able to differentiate into all cell types of the body. *Adult stem cells* are pluripotent undifferentiated cells found throughout the body after embryonic development. Unlike embryonic stem cells, adult stem cells can only form a limited set of cell types and function to replenish dying cells and regenerate damaged tissues. Given the importance of driver cells for development in ET, a spontaneous question regards the relation between driver cells and stem cells. The definition reported above hides the fact that the term "stem cell" can be used in biology to refer either to i) cells that are capable of self-renewal, or ii) cells which are very plastic and can be induced to turn into a specific type. In adherence to this distinction, we have chosen to keep the name "stem cell" for ii) and to use the name "driver cells" for i).

Many analogies exist between the concept of driver cell and the concept of Spemann's organiser, which is an area of the Xenopus embryo able to induce embryonic primordia upon transplantation into a different location [3]. Likewise, if a driver cell destined to give rise to a part of the structure is moved to a different position, that structure part will grow in the new, ectopic position. In our model a single driver cell can produce a body structure in a fully autonomous way. When proliferating, such driver cell will itself induce the creation of new driver cells which, in turn, will become the centres of other proliferation or apoptosis events, from which other driver cells will be created and so on, until the whole structure is generated. This appears to be consistent with the so-called "head / trunk / tail organiser model" [28], which foresees the presence of many organisers. In our model the organisers (the driver cells) are many, are hierarchically structured and are continuously created during development.

Based on these considerations, we propose a new classification of biological cells (Fig. 8):

1. driver cells. These are the natural counterpart of driver cells in the model and correspond to biological organisers. They can create and / or induce stem cells to undertake actions and commit to specific fates;

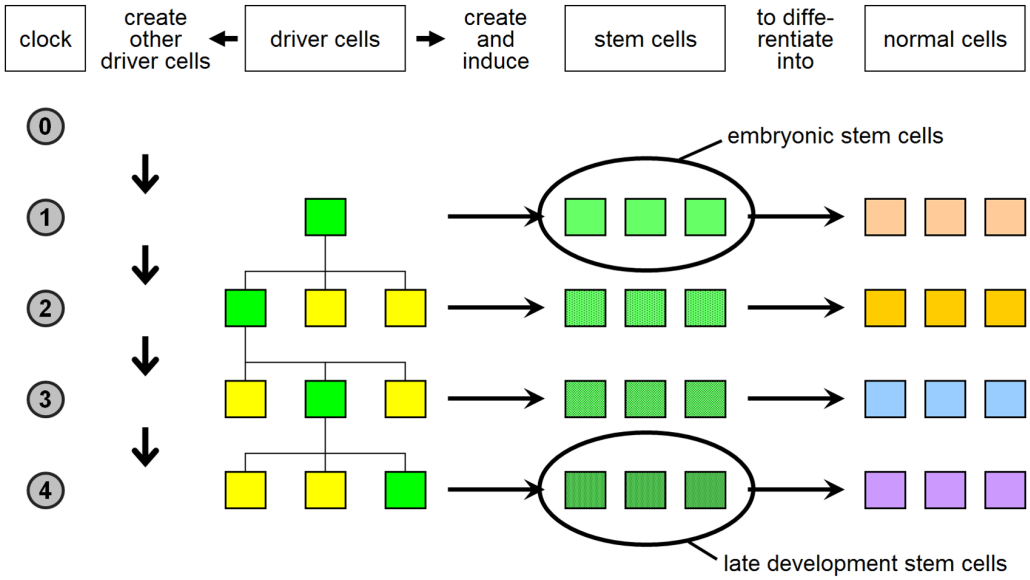

Figure 8: Proposed classification of biological cells. Driver cells marked in green become active during development. As a result, they create and / or induce stem cells to differentiate into normal cells. In parallel, other driver cells are created. Normal cells can be induced to further differentiate, or trans-differentiate, or even de-differentiate by other driver cells at a subsequent stage. "Darker" stem cells have a lower degree of plasticity.

2. stem cells. These are cells characterised by a high degree of plasticity and susceptibile to be induced by driver cells to undertake actions and commit to specific cellular fates;

3. normal cells. These are cells with no plasticity, i.e. they are terminally differentiated cells. Actually we can imagine stem cells and normal cells as the extremes of a continuum of cells characterised by a decreasing degree of plasticity.

The innovation in this scheme is the introduction, besides normal and stem cells, of a third layer of cells, namely driver cells, which represent the scaffold of the organism. During development some of them are activated thanks to specific developmental genes. As a result of their activation, a local pool of stem cells, specific for the body part under construction, is produced and induced to differentiate into the specialised cell types needed. At the same time, other driver cells are generated and homogeneously distributed in the part just created, some of which are destined to become centres of other developmental acts in the course of development. The plasticity of stem cells depends on the driver cells that generates them and, in general, we can think of it as a decreasing function of developmental time.

As the example of organ development reported in Fig. 7 shows, the set of MOC values generated during development has a tree structure. This descends from the fact that each driver cell originates from the conversion of a normal cell, which has a single mother. This properties is not only true for parts of the structure, but also applies to the structure as a whole. This appears to be coherent with the information on the set of stem cells involved in the generation of particular organs or systems, such as the hematopoietic system [25]. Our model provides also a means to bridge the conceptual gap between embryonic and adult stem cells. In our model embryonic driver cells (and relevant stem cells) correspond to elements in the tree near the root, while adult driver cells (and relevant stem cells) correspond to the leaves of the tree.

3 Evo-devo and transposable elements

3.1 The evo-devo process

Given a certain genome embedded in a single cell put on the grid, we have now a set of rules to generate a development and obtain a final structure. A naturally arising question is: how do we choose the genome in order to produce a predefined target structure? The answer is: we use a *Genetic Algorithm (GA)*, a technique that simulates Darwinian evolution. The GA evolves a *population* of genomes, each of which guides the development of a structure starting from a single cell initially present on the grid, for a number of generations. The members of the population are called *individuals*, a term that we will use as a synonym for structure in an evolutionary perspective.

At each generation, all genomes in the population (one at a time) guide the development of the structure from the zygote stage to the final phenotype, whose adherence to a target structure is employed as a fitness measure. This operation is repeated for all genomes, so that eventually each genome is assigned a fitness value: based on this value the genomes are then selected and randomly mutated, to produce the new population. This cycle is repeated until a satisfactory level of fitness is reached.

The coupling of the model of cellular development and the GA gives origin to an evo-devo process to generate 3-dimensional structures. "In silico" experiments (i.e. computer simulations, see examples in Figs. 9, 10) have proved the effectiveness of the process in "devo-evolving" structures of any kind and complexity (in terms of e.g. number of cells, number of colours, etc.). The power of this method essentially depends on the features of the model of development, in particular on the presence of a homogeneous distribution of driver cells, which keeps the structure "plastic" throughout development. On the other hand, the speed of the evolutionary process is also safeguarded by a special procedure, which will be described next.

3.2 Germline Penetration

In ET most driver cells produced during development do not orchestrate any events (are inactive). This may correspond to a driver cell which has a MOC that matches a MOS in the right part of some developmental gene, but the TM in this right part has a value higher than the maximum number of developmental stages, or the driver is produced at a stage which is "later" than any of these TM, or all the genes with a matching MOS have the switch set to inactive. A simple germline mutation may change the situation.

It is also possible that some driver cells are inactive because there is no MOS in the genome which would match their MOC (Fig. 11). In biological terms it may mean that in this particular state of the driver, all the regulatory factors (proteins, RNA) present in the cell do not result in the activation of all the genes which are necessary to allow for an event to happen. In ET the MOC and the MOS are encoded as long sequences of numbers. Therefore, the probability that a suitable MOS emerges in the genome simply through mutations and recombinations is very low. A countermeasure consists in "suggesting" to the GA MOS sequences likely to match existing MOC values. If we suggest the GA to use as genes' MOS sequences some of the MOC values generated during the development of the structure, the match is guaranteed.

This idea is implemented in a procedure called *Germline Penetration*, executed at the end of the development of each individual structure. Germline Penetration copies at random (some) MOC values generated during the development of the structure onto MOS sequences of developmental genes contained in a special copy of the genome called "germline" genome, distinct from the genome incorporated in all cells. The germline genome, after reproduction

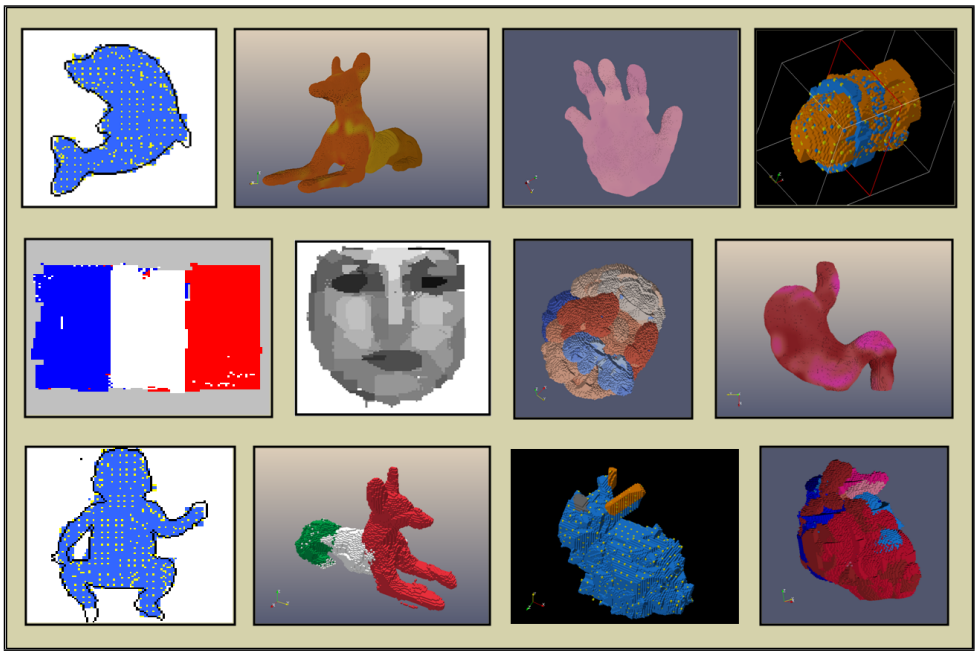

Figure 9: Examples of structures generated using ET. All panels show the final stage of development of the best evolved individual.

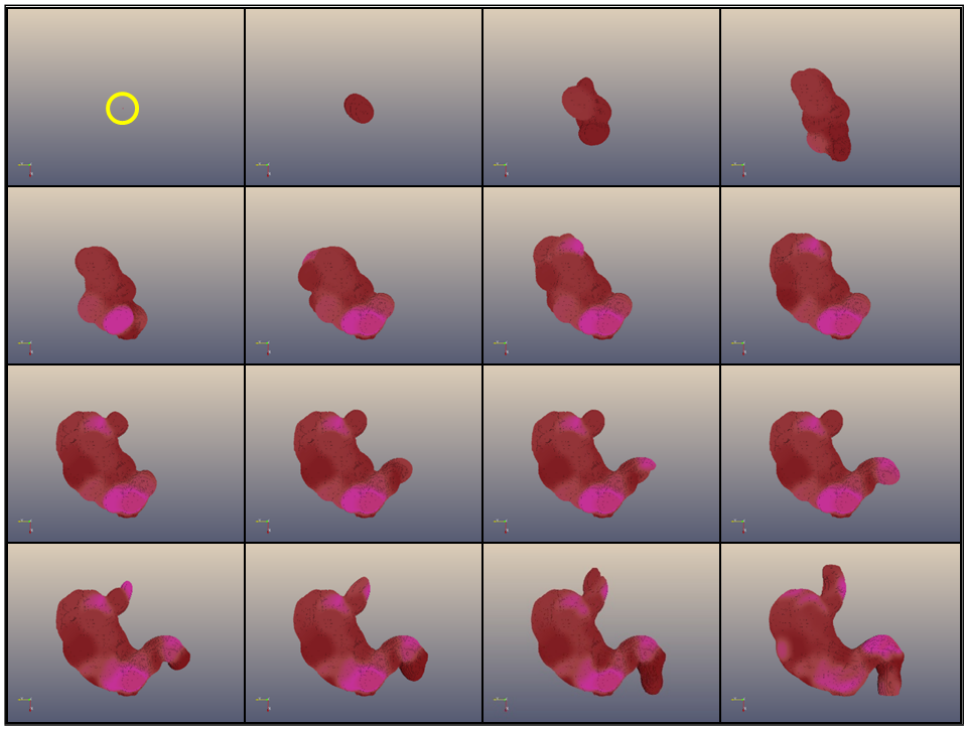

Figure 10: Development of a 3-dimensional coloured structure representing a human stomach obtained with ET.

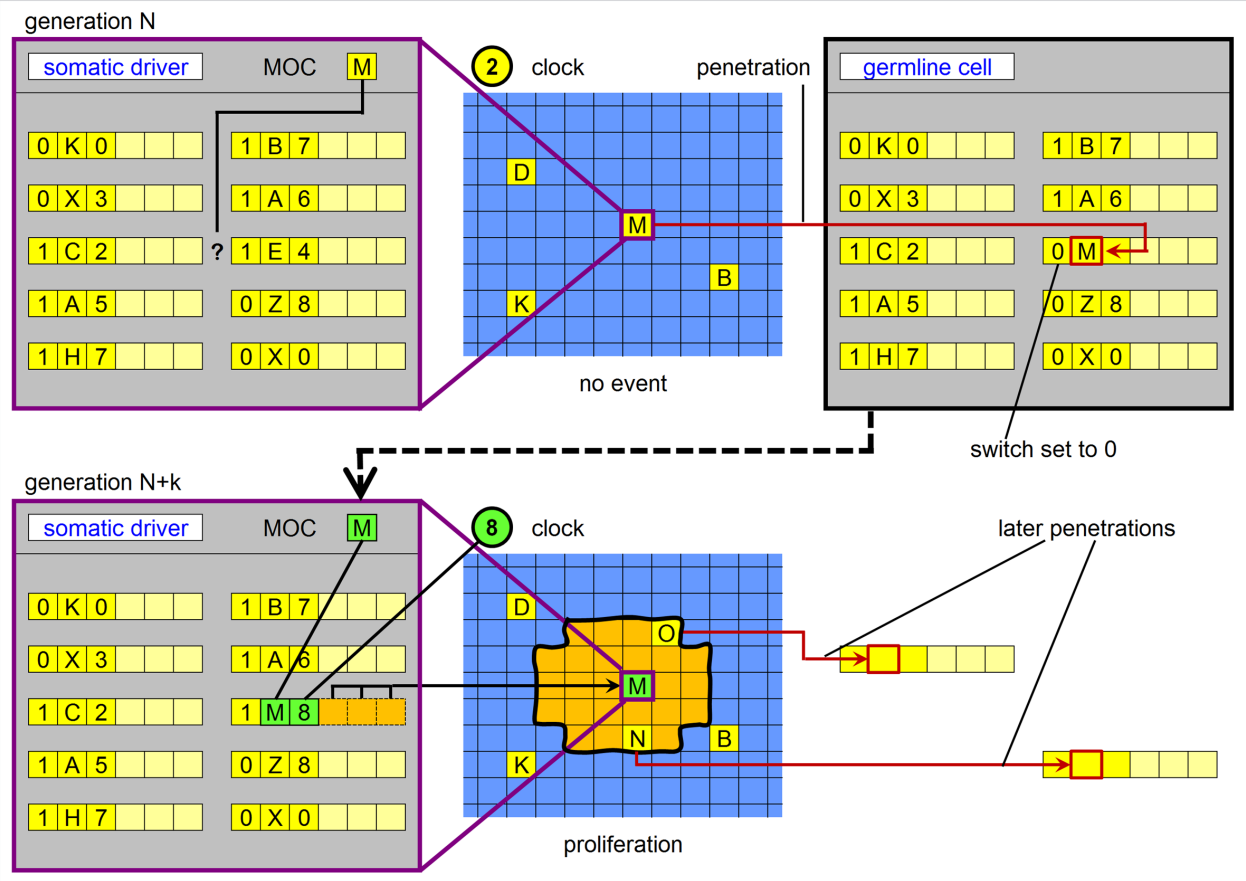

Figure 11: Germline penetration of new regulatory sequences. In the upper part, in a driver cell the genome is not able to "respond" to the current cellular state M with a corresponding event. A new MOS able to respond is created in the driver cell. The new MOS leaves the driver cell and reaches the germline, where it is incorporated in the genome. In a subsequent generation, when a cell reaches cellular state M, its genome (inherited from the penetrated germline) can respond with a specific genetic unit. The result is a proliferation event.

and the application of mutations and recombinations, is destined to become the ("somatic") genome of the individuals of the next generation, in which it will again be embedded in all cells.

Fig. 12 shows the interplay between development and evolution, mediated by Germline Penetration across multiple generations. Once evolution is provided with "good" genes' left parts, i.e. left parts containing MOS sequences that are guaranteed to match existing MOC values, it has to optimise the relevant right parts, a process that can take several generations. When the optimisation of the right part is completed, the new developmental genes can be activated and carry out as many change events. The new MOC values generated as a result of the new change events occurred are again transferred to the germline genome to be embedded into the genome of the offspring and the whole cycle repeats itself.

The genes with the copied MOS sequences are initially set as inactive, as they would otherwise all become active with a non-optimised right part, causing a major disruption in development (and an abrupt decrease in the individual's fitness). Their activation, obtained through a "flip" of the switch, is left to a subsequent genomic mutation. As a consequence, at any given time in the course of evolution of any individual, most developmental genes in the genome are inactive. Germline Penetration is essential in ET for the evolution of complex

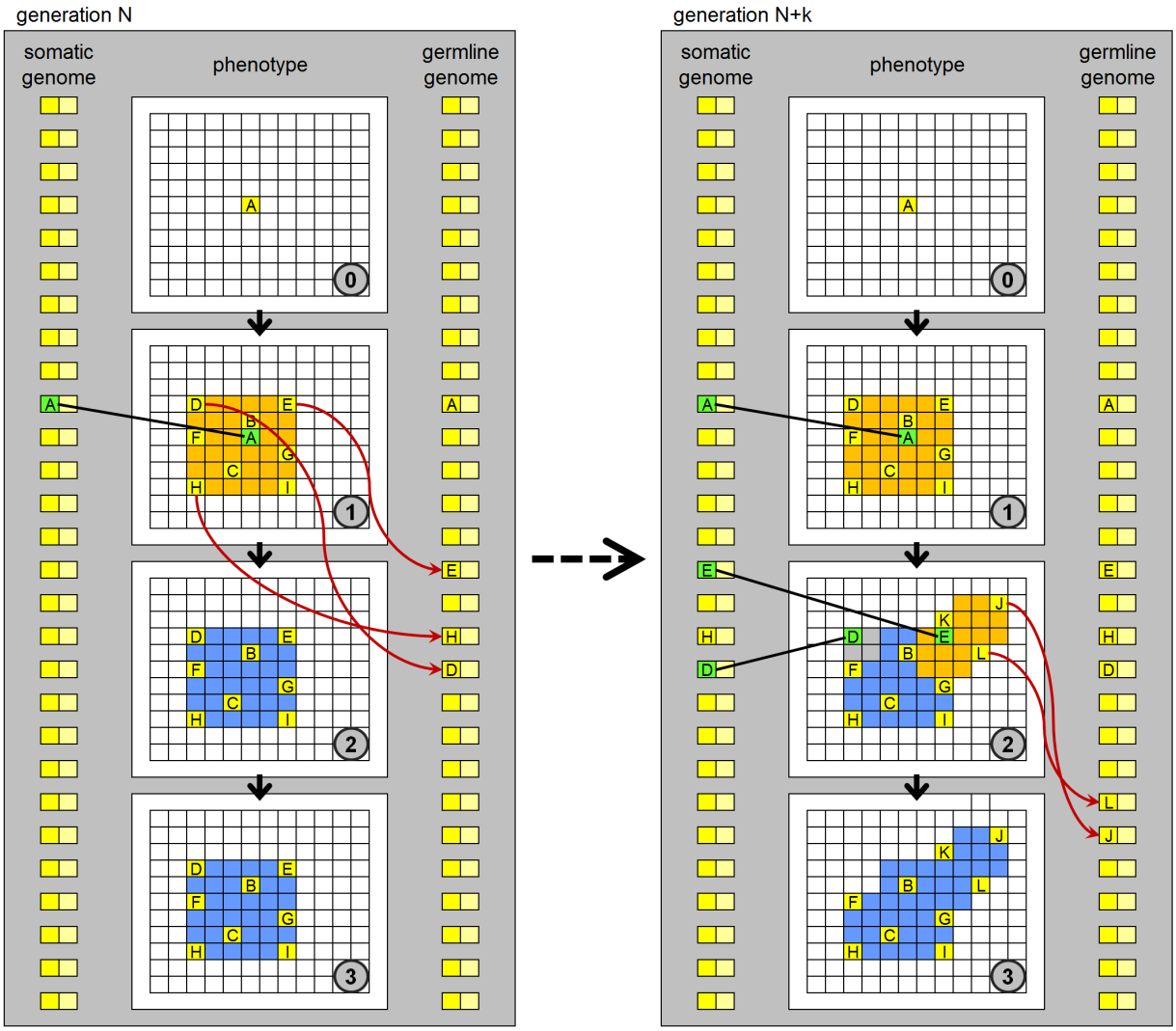

Figure 12: Germline penetration shaping the genome. On the left: development of an individual at generation N. Development stops at stage 1. MOS values matching some of the MOC values generated during development leave the respective driver cells and are transferred to the germline genome. The germline genome is inherited by the individuals in the next generation. On the right: development of an individual at generation $N + k$. New genes, derived from the penetrated elements, are present in the genome, whose MOS values match some of the MOC values generated in stage 1, so development can proceed further than at generation k. The new MOS values generated in stage 2 can again penetrate to the germline genome.

shapes: our experiments in silico demonstrate that when Germline Penetration is disabled, the evolutionary process practically grinds to a halt.

3.3 Transposonal theory of heredity

According to ET, the cycle shown in Fig. 12 represents the evo-devo core of multicellular life. The central role played by Germline Penetration in our model lead us to hypothesise the existence of a similar procedure also in biological systems. For the implementation of a biological Germline Penetration, we need genetic elements able to build new regulatory sequences in somatic cells, and susceptible of being transferred to germline cells. Transposable elements, first discovered by B. McClintock [21], seem to possess the characteristics necessary for this role.

Transposable elements (TE), or transposons, are DNA sequences that can move around to different positions in the genome. During this process, they can cause mutations, chromosomal rearrangements and lead to an increase in genome size. Major subclasses of transposons are represented by DNA transposons, LTR retrotransposons, long interspersed nuclear elements (LINEs), and short interspersed nuclear elements (SINEs). Despite representing a large genomic fraction (30-40% in mammals), no clear function has yet been identified for transposons, which have therefore been labelled as "junk DNA", until recent research suggested they could indeed have a biological role. TEs have been studied along two main dimensions: evolution and development.

A first line of research has been concerned with the dynamics of TE diffusion in genomes across multiple generations, as can be evidenced through modern genome-wide analysis techniques. Transposable elements appear to be temporally associated with major evolutionary changes [11] [23] and many elements are present only in specific lineages (Alu in primates for example), suggesting a causal link between the appearance of such elements in the genome and the evolutionary change that originated the lineage. Lineages whose genomes are not subject to intermittent TE infiltrations are more static from an evolutionary point of view. A specific increased activity of some TE families in the germline has been observed [22], in agreement with the hypothesised evolutionary role of transposons. As a matter of fact, only transposons that become fixated in the germline can be passed on to the next generation and have trans-generational effects.

TEs were initially characterised as elements controlling phenotypic characteristics during development in maize [21], when they can insert themselves near genes regulating pigment production in specific cells, inhibiting their action and making the cells unable to produce the pigment. This regulatory role is supported by further research indicating that (in human) transposons tend to be located near developmental genes [20]. Recent observations corroborate the view that TEs are active in somatic cells, opening the possibility that they can bring diversity among somatic cells with the same genome [2]. All together this evidence suggests that TE activity can be grouped in two categories: i) activity in somatic cells during development and ii) activity in germline cells and in early development.

Based on these considerations, we can hypothesise a mechanism for the implementation of a biological Germline Penetration. Our hypothesis is that, in a driver cell not able to produce a change event because no biological MOS is able to match the biological MOC, TEs become mobilised and start "jumping" around the genome. This process can lead to the creation of many new regulatory sequences, hugely increasing the chances that a lucky combination of such sequences be able to match the biological MOC. When this occurs, a new change event can take place in the somatic cell. We further hypothesise that, when this occurs, the TEs which helped to build the new sequences exit the cell and make their way into a blood vessel.

Through the bloodstream they are carried to germline cells, where they insert themselves into their genome, thus allowing the innovation to be conveyed to the next generation.

When the TEs reach the germline, they can insert themselves into the same genomic positions as in the somatic driver cell from which they originated. This requires from TEs the ability to choose with great precision their insertion site. Alternatively, the TEs could be inserted in the genome in random positions. The coding regions falling under the regulatory domains of the inserted sequences would then come under selective pressure and become a biological right part of a biological developmental gene. After a number of generations evolution would find good solutions for such right parts, allowing development to move ahead.

The mechanism described has interesting evolutionary implications. Such implications can be deducted considering that in ET, whenever a driver cell is triggered to proliferate by a developmental gene, a "wave" of new driver cells, each with its own MOC value, is created in the body of the (new) species. The action of Germline Penetration translates this wave of new MOC values in somatic cells into a corresponding wave of new MOS sequences spreading in the germline genome and subsequently in the somatic genome of future generations. Such events during evolution coincide with moments in which major changes occur to the evolving species, causing new body parts or features to appear. In biological terms, this means that the spreading in the genome of waves of new sets of TEs in the course of evolution corresponds to moments in which new branches (new species) are generated in the "tree of life". Such predictions made with our model on purely theoretical grounds, appear to be confirmed by experimental evidence [23].

More precisely, our model predicts the precise following sequence of events associated to a change in the lineage (in biological terms):

1. a driver cell orchestrates a change event, which results in the creation of a new body part in a species (and hence a change in the lineage and creation of a new species);

2. new driver cells and relevant transcription factors are created in that body part (wave of new transcription factors); these factors could give origin to other change events, but the genome lacks regulatory sequences able to match;

3. matching regulatory sequences are created through TEs, enabling the genome to respond to the new transcription factors;

4. Germline Penetration "pumps" the matching regulatory sequences from the somatic cell to the germline genome (wave of new transposons in the genome);

5. Darwinian evolution can now work on the coding sequences regulated by the penetrated regulatory sequences and the cycle starts over from 1.

According to ET this is the core evo-devo cycle, and the engine of multicellular evolution, which is essentially Lamarckian for regulatory sequences and Darwinian for coding sequences. One point is worth stressing. The experimental evidence [23] suggests that the spreading of new transposon families in the genome and the occurrence of major changes in the relevant lineage are simultaneous events. This seems to hint that the colonisation of the genome by the transposons was the driving force behind the change in the lineage. Our interpretation of this phenomenon is different. In fact, according to ET, the spread of new transposon families in the genome is an event which comes immediately (in evolutionary terms) *after* the change, not before. The confirmation of this prediction, made possible by techniques to estimate the age of DNA sequences more sensitive than those currently employed, would represent a clear indication in favour of the model.

Sperm-mediated gene transfer (SMGT) [27] is a procedure through which new genetic traits are introduced in animals by exploiting the ability of spermatozoa to take up exogenous DNA molecules and deliver them to oocytes at fertilisation. The reverse-transcribed molecules are propagated in tissues as low copy extrachromosomal structures, able to inducing phenotypic variations in positive tissues, and transferred from one generation to the next in a non-Mendelian fashion. Experimental evidence suggests that sperm-mediated gene transfer is a retrotransposon-mediated phenomenon. The ability of spermatozoa to take up DNA molecules and incorporate them into their genome could represent the mechanism used by nature to implement the final stage of the biological Germline Penetration, when the new regulatory sequences enter the germline genome. ET provides the biological meaning of the properties which make SMGT possible.

In conclusion, in our model Germline Penetration represents an indispensable tool to boost evolution. This point deserves emphasis: in ET, for multicellular structures, Darwinian evolution *is not sufficient*. In this perspective development and evolution appear to be two sides of the same process (linked together by Germline Penetration): the sentence "nothing in biology makes sense except in the light of evolution" can be reformulated as "nothing in multicellular biology makes sense except in the light of devo-evolution".

4 Junk DNA and ageing

4.1 Facts and theories on junk DNA and ageing

In biology the term "junk DNA" is used to label portions of the genome which have no function or for which no function has yet been identified. Many junk sequences appear to have been conserved over many millions of years of evolution, which seems to hint that they play an essential role: eukaryotes appear indeed to require a minimum amount of junk DNA in their genomes. In the human genome, as far as 95% of the genome can be designated as "junk". Major categories of junk DNA are represented by i) introns, non-coding sequences within genes; ii) chromosomal regions composed of residues of once functional copies of genes, known as pseudogenes; iii) transposable-elements. This last category alone represents some 30-40% of the genome of mammals.

A number of hypotheses have been proposed to explain the presence of junk DNA in the genome of many species. Junk DNA could represent a reservoir of material from which new genes could be selected: as such, it could be a useful tool used by evolution. Some junk sequences could be spacer material that allows enzymes to bind to functional elements. Some junk DNA can serve the purpose of preserving the integrity of the cell nucleus from a mechanical / structural viewpoint. Finally, junk DNA could be involved in regulating the expression of protein-coding genes. What is clear is that, while the amount of coding DNA appears to be similar across a wide range of species, the amount of junk DNA displays a much broader range of variation and seems indeed to be correlated with organismal complexity [13].

Ageing is a process that occurs in the lifespan of most living beings. It involves the accumulation of changes in the organism over time, leading to a progressive deterioration of bodily functions. Ageing does not affect all species: in some simple species, its effects are negligible or undetectable. Moreover, the rate of the process is very different among affected species: a mouse is old at three years of age, a human at eighty years of age. Differences exist also between genders (women have on average a longer lifespan compared to men) and among single individuals. Both genetic and environmental factors seem to be involved in the ageing phenomenon. Many genes have been identified, that increase the lifespan in some species; a measure that was proved effective in reducing the rate of ageing across many species is caloric restriction.

Many theories of ageing exist. **Stochastic theories** blame environmental factors that induce damage on living organisms as the cause of ageing. Within this category, the *wear-and-tear theory* argues that ageing is the result of damage accumulating over time (like the "ageing" of a mechanical device). The *free-radical theory* is based on the idea that free radicals induce damages in cells, which at the organismal level produce the ageing phenotype. According to **programmed theories** ageing is regulated by biological clocks, which trigger changes to maintenance and repair systems. Within this category, the *ageing-clock theory* maintains that ageing results from a programmed sequence, triggered by a clock built into the functioning of the nervous and / or endocrine systems. The shortening of the telomeres at each cell division could represent a possible physical implementation of the clock.

Evolutionary theories [12] consider evolutionary phenomena as the main cause for the differences in the ageing rates observed across different species. The *mutation accumulation theory* states that gene mutations affecting individuals at a young age (before reproduction) are strongly selected against, while mutations to genes displaying their effect at an old age experience no selection, because the individuals have already passed their genes to the offspring. The *antagonistic pleiotropy theory* argues that strategies leading to a higher reproductive fitness and a shorter lifespan would be favoured by natural selection. The *disposable soma theory* claims that mutations which save energy for reproduction (positive effect), by

disabling molecular proofreading devices in somatic cells (negative effect), would be favoured be natural selection. Ageing would be the cumulative result of the negative effects on the organism.

4.2 Interpretation of junk DNA and ageing

In the ET framework, for any individual the set of MOC values (SMV) generated during development can be divided into i) MOC values that activate a developmental gene during development and ii) MOC values that do not activate any developmental gene during development. Likewise, the set of developmental genes (SDG) can be divided into i) developmental genes that are activated during development and ii) developmental genes that are not activated during development. By analogy with biological systems, elements in the two categories labelled with ii) (inactive elements) can be defined as "junk" MOC values and "junk" developmental genes. Typically, sets labelled with ii) are much larger than sets indicated with i): in our experiments the average ratio between the number of MOC values used and the total number of MOC values generated is around 10%.

Indeed, the ET machine cannot do its job without generating a lot of junk MOC values. The only way to avoid it would consist in reducing the density of driver cells (system parameter), but this would also reduce the effectiveness of the morphogenetic process. Therefore, the presence of a certain amount of junk MOC values is unavoidable. On this material Germline Penetration acts like a shuttle, transferring junk MOC values onto a corresponding number of junk MOS sequences in the genome. The developmental genes in which the insertions take places are junk genes because their switches are initially set to OFF. Without this measure, they would all become active at once with a non-optimised right part, causing disruptions in the development of the individual.

Therefore, the presence of junk information in both the set of MOC values and the genome is a fact which is inescapably connected to the core of the ET machine, a requirement essential to its *evolvability*. Recalling how MOS sequences are hypothesised to be implemented in nature as sets of transposable elements, we can conclude that our model provides an explanation for the presence of this category of junk DNA which, as we have seen, accounts for up to 40% of the genomic content in mammals. As it will become clear, the presence of MOC values and genes not activated in the course of development plays a role also in ageing phenomenon.

For a given individual development unfolds in N developmental stages; at the end of it the fitness is evaluated, and the genome content is passed over to the subsequent generation. The moment of fitness evaluation, that in nature can be thought to roughly correspond to the moment of reproduction, in our experiments has always coincided with the end of the simulation. On the other hand, we can imagine to let the global clock tick on and see what happens in the period after fitness evaluation. The distinction between the periods before and after fitness evaluation correspond to the biological periods of development (say, until 25 years of age in humans -the average age of reproduction) and ageing (from 25 years of age onwards).

Fig. 13 shows how the situation looks like at the end of an individual's development. Many driver cells are present in the body of the individual, which have not been activated during development. This stock of junk driver cells represents a reservoir of change events that can potentially occur in the period after fitness evaluation, when they are (by definition) not affecting the fitness value. For this reason the developmental genes which carry out the events experienced no selective pressure in the course of evolution, and have non-optimised genetic sequences: as a result these events will tend to have a random-like nature. In fact, they are *not* random, as they are encoded in developmental genes. Their nature can be best characterised as "pseudo-random", not unlike computer-generated sequences of random numbers.

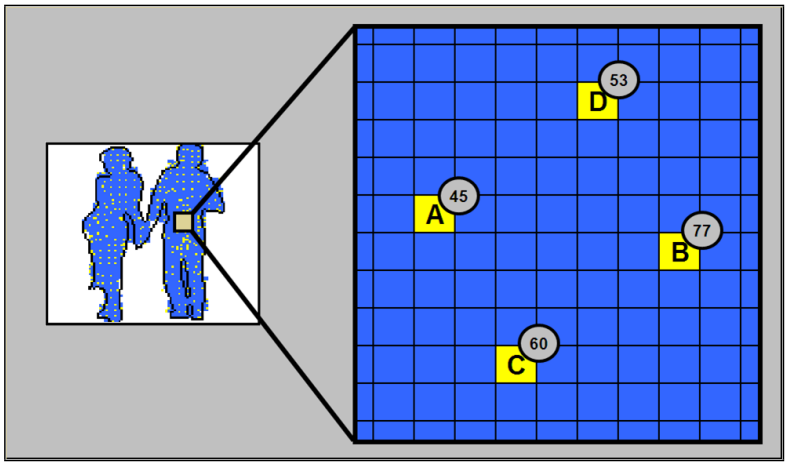

Figure 13: Ageing driver cells at the end of development. The timer value of the developmental gene which is going to be activated is indicated in the circle (numbers refer to age in humans expressed in years).

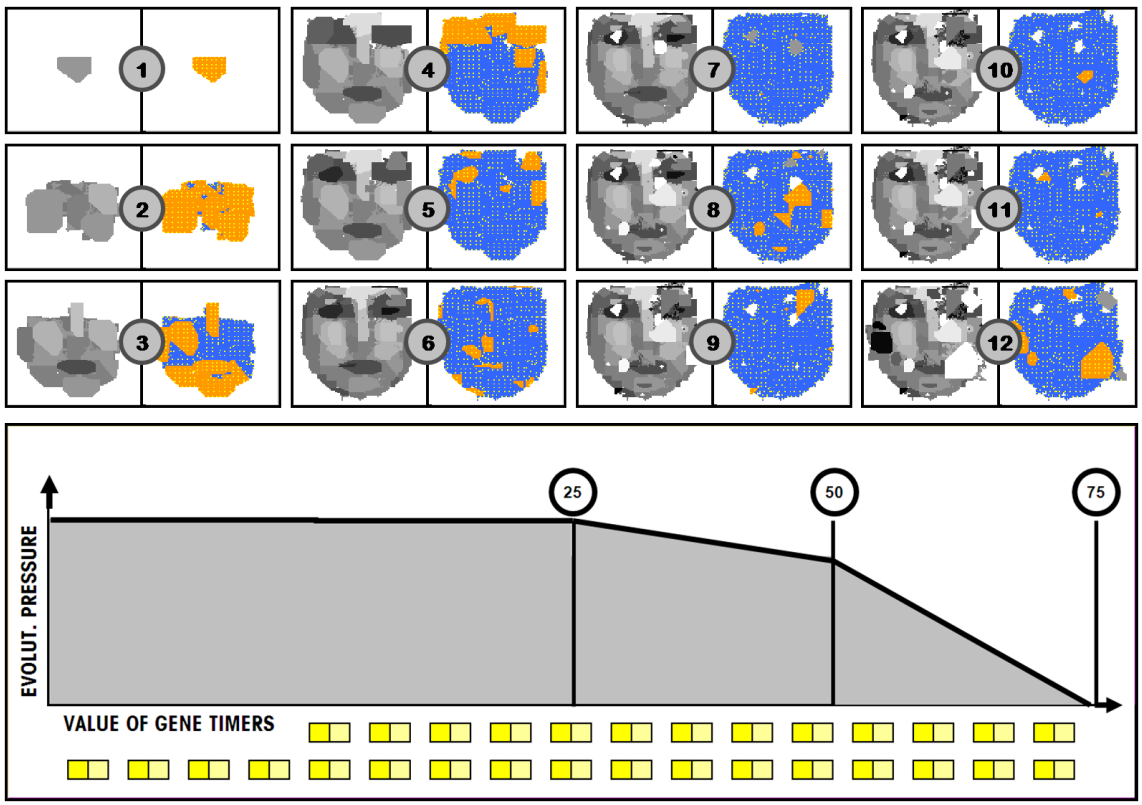

Figure 14: Simulation of ageing for a face. In the upper part: simulation of an ageing face. On the left the period of development (stages 1-6): the structure grows from a single cell to the mature phenotype in stage 6, when fitness is evaluated. On the right the period of ageing (stages 7-12): the picture quality (its "health") deteriorates steadily under the action of non-optimised developmental genes. In the lower part: the evolutionary pressure acting upon a development gene is an inverse function of the value of the gene's timer. Genes with timer values (referred to human life) lower than 25 years are subject to high evolutionary pressure, genes with timer values greater than 25 years are subject to a steadily decreasing pressure, until their effects become pseudo-random.

Such sequences look erratic, and yet successive runs of the programme produce exactly the same sequence, over and over again.

The consequence of the pseudo-randomness of change events occurring after fitness evaluation is that their effects on the overall individual's health are more likely to be detrimental than beneficial. Here, "fitness" has to be interpreted as the *reproductive* fitness of the individual (i.e. the probability that its genes will survive in the future genetic pool of the population), while "health" represents the individual's "physical" condition. Therefore, in the ageing period, the individual's health will tend to progressively decrease over time under the action of pseudo-random events: this, we think, is the deep nature of the ageing phenomenon. In other words, ageing can be seen as a *continuation of development*, driven by developmental genes activated in specific driver cells after fitness evaluation.

Based on these considerations the set of all MOC values generated during an individual's development can be divided into: i) set of MOC values active during development; ii) set of MOC values active during ageing; iii) set of MOC values never active. Analogous distinction can be done also for the set of developmental genes, into: i) set of developmental genes active during development; ii) set of developmental genes active during ageing; iii) set of developmental genes never active.

The hypothesis according to which the evolutionary pressure acting on a developmental gene is zero if the gene becomes active after reproduction is rather simplistic. In fact, in nature, the reproductive fitness of an individual also depends on events manifesting themselves after reproduction, which also affect the survival chances of the progeny. In other words the effect of an event on the fitness (and hence the evolutionary pressure acting upon the corresponding gene) tends to decrease as the age of its appearance (determined by the gene timer value) increases, rather than vanishing right after reproduction.

Fig. 14 reports a computer simulation of ageing for a "face", based on the gradual reduction of selective pressure. The lower part of the Figure shows the curve of the pressure as a function of the timer value of genes. The pressure is high until the age of 25 years, when reproduction occurs; then it starts declining with a mild slope, since the individual has to be in good shape to look after its children, who are unable to survive alone. At the age of 50 years, the slope of decline becomes steeper, as its parental role is less important and the ex-children have children themselves. Until the age of 75 years, the individual can play a role as a grandparent, but after this moment whatever happens to our individual has no influence on the survival of its progeny: therefore, genes activated in this age range are subject to no evolutionary pressure.

The view of the ageing process as a progressive accumulation of pseudo-random events having little or no effect on reproductive fitness provides a new interpretation for the genetic diseases typically associated to the old age. Examples of such diseases are Alzheimer disease, type II diabetes, heart problems, and in general all diseases that seem to have an idiopathic or intrinsic origin (i.e. they are not caused by external agents such as viruses or bacteria) but whose onset in humans typically occurs from the 5th decade of life onwards. These diseases seem indeed to be caused by the malfunctioning of specific genes and could therefore be labelled as genetic diseases, even though they are not present at birth but manifest themselves only later in life.

The hypothesis we are proposing here is that the difference between the phenotypic manifestations of such diseases and the effects of "normal", "healthy" ageing are rather quantitative than qualitative. Both normal ageing and ageing-associated diseases are indeed driven by change events caused by developmental genes whose timers are set to go off in the old age: the manifestations related to normal ageing are only milder, more benign than those associated to the diseases. This interpretation provides a straightforward explanation on why the

temporal patterns of ageing and ageing-related diseases are coincident: they are basically the same thing.

As we mentioned, a slow-down in the pace of ageing can be obtained through caloric restriction. One possible explanation of this experimental evidence is that caloric restriction directly affects the functioning of the global clock, which becomes slower. As a result, all developmental genes set to go off in the ageing period are delayed, all by the same amount. Some genes, which are known to slow-down ageing in some species, could exert an influence on the cellular molecular machinery which transduces the clock (likely implemented as a protein) into the nucleus. In this way the value of the clock perceived inside the cell will be lower.

The theory proposed represents a synthesis of elements coming from all major categories of ageing theories. While stochastic theories blame random events accumulating over time as the cause of ageing, programmed theories see ageing as the effect of a programme unfolding according to a biological clock. These two explanations appear mutually exclusive, and yet evidence exists in favour of both. ET provides the synthesis through the concept of pseudo-randomness: a programme composed of pseudo-random events is both deterministic and random-looking. The model proposed fits also in the class of evolutionary theories and is basically consistent with the mutation accumulation theory.

We wish to conclude this section dedicating a final comment to the role played by elements inactive during development. We have shown how a part of these elements is actually devoted to cause pseudo-random events in the ageing period, relegating to the inactive elements the role of true junk. On the other hand, elements are free to move between sets: the sets of inactive elements represent therefore a reservoir of potential new events, and a tool to explore new developmental trajectories. These considerations bring us to deducting a direct link between the evolvability of a species and its susceptibility to ageing, being both aspects mediated by the presence of a big stock of junk. The fact that bats have unusually small genomes (i.e. little junk) and display a remarkably long lifespan (i.e. they appear to age less) among mammals of comparable dimension [30], appears to be consistent with this hypothesis.

5 Cancer

5.1 Teratomas

In this section we will analyse a possible malfunction of the model and we will show how such a malfunction gives origin to a phenomenon that mimics a particular kind of tumour called a teratoma. A teratoma is a tumour with tissue or organ components resembling normal derivatives of all three germ layers. The tissues of a teratoma, although normal in themselves, may be quite different from surrounding tissues, and may be highly inappropriate, even grotesque: teratomas have been reported to contain hair, teeth, bone and very rarely more complex organs; usually, however, a teratoma does not contain organs but rather tissues normally found in organs such as the brain, liver, and lung. Teratomas are thought to be present at birth, but small ones often are only discovered much later in life.

In our model, a certain body part of an organism can be generated by a single driver cell that, once activated, proliferates. From this proliferation other driver cells are created, some of which get in turn activated and proliferate, leading to the generation of other driver cells, etc. (the same holds true for the entire organism). This process, shaped by evolution to occur in a precisely orchestrated fashion, presupposes that each driver cell, at the moment of activation, finds itself in the right position, surrounded by the right cellular micro-environment: only in this case is the cascade of events originated from the driver cell's activation capable, along with physics, of generating the relevant body part.

This delicate mechanism can be perturbed in many ways. We will now focus on a case characterised by a mutation that, at stage S_j, turns the MOC value M_j of a certain driver cell D_j, positioned at point P_j, into another MOC value (M_k). If the MOC value M_k is not generated during normal development, or if it is generated but never activated, nothing happens. The situation is different if the MOC value M_k does become active during normal development to produce a certain body part, say at stage S_k, when cell D_k finds itself at point P_k. In this case, as a result of the mutation, the cascade of events destined to give rise to such body part will start from both point P_k at stage S_k (right place and moment) and point P_j at stage S_j (ectopic place, wrong moment). Being activated in the wrong place and moment, cell D_j is not surrounded by the right micro-environment: therefore, the cascade of events originating from D_j will only manage to mimic the development of the relevant body part in a grotesque fashion.

For example, the biological MOC of the driver cell destined to give origin to the lung could be turned during embryonic development into the biological MOC of the zygote. As a result, the development of the whole embryo would start over again from this cell: the cell would proliferate, generating other MOC values some of which, as in normal development, trigger other proliferation events, etc. But, since in this case the zygote and all other MOC values originated from it would be in ectopic positions and surrounded by wrong micro-environments, while the different cell types would continue to be created, the interactions with other cells would prevent them from being arranged in the correct patterns. Instead, an amorphous mass of differentiated cells would be produced. We will now turn our attention to carcinogenesis in the general case.

5.2 Facts and theories on carcinogenesis

Carcinogenesis is the process by which normal cells are transformed into cancer cells. Based on studies of skin cancer, the process of carcinogenesis is traditionally divided into three phases: initiation (linked to chemicals or physical insults that induce permanent alterations to DNA), promotion (the proliferation of the initiated cell induced by subsequent stimuli) and progression (the stepwise transformation of a benign tumour into a malignant one). The

"standard theory", also referred to as the "multi-hit" hypothesis [17], states that carcinogenesis is a multi-step process that can take place in any cell, driven by damage to genes that normally regulate cell proliferation. This upsets the normal balance between cell proliferation and cell death and results in uncontrolled cell division and tumour formation. A critical point inherent to the standard theory is that not all cells of a tumour seem able to induce a secondary tumour when injected into nude mice [31].

This latter piece of evidence is addressed by a more recent theory [1] which traces back the origin, the maintenance and the spread of a tumour to a relatively small subpopulation of cells called **cancer stem cells (CSCs)**, whereas the bulk of the tumour would actually be composed of non-tumorigenic cells that, deprived of the cancer stem cells, would quickly shrink and disappear. CSCs possess characteristics associated with normal stem cells, specifically the ability to give rise to all cell types found in a particular cancer sample; CSCs may generate tumours through the stem cell processes of self-renewal and differentiation into multiple cell types. The implications of this hypothesis for therapy cannot be overstated: conventional chemotherapies kill differentiated or differentiating cells, which form the bulk of the tumour but are unable to generate new cells; a population of CSCs, which gave rise to it, could remain untouched and cause a relapse of the disease.

Since cancer is a disease of genes, the attempt to link cancer or specific cancers to patterns of gene mutations, consistently found in all tumour samples, would seem logical and well-founded. A few cancer-related genes, such as p53, do seem to be mutated in the majority of tumours, but many other cancer genes are changed in only a small fraction of cancer types, a minority of patients, or a subset of cells within a tumour. Although the effort to reconduct tumour formation to subsets of mutated genes in 100% of cases has so far been unsuccessful, it is nonetheless undeniable that *correlations* between different tumours and specific patterns of mutations exist, i.e. individual genes are mutated in percentages that are tumour-specific [32]: these correlations represent evidence a theory of carcinogenesis has to explain.

The presence in tumours of cells of different types and / or having different degrees of differentiation is a well documented phenomenon, coherent with the cancer stem cell theory. There is increasing evidence that diverse solid tumours are organised in a hierarchical fashion and their growth is sustained by a distinct subpopulation of CSCs. This fact is difficult to explain by the standard theory, which postulates that all tumour cells are derived clonally from a single cell, the first to have accumulated the number of hits required to achieve the malignant transformation. According to the standard theory, genotypic and phenotypic diversity within a tumour can only be realised through the effect of subsequent mutations. Evidence, on the other hand, points to the existence of a structured hierarchy of cell types, displaying variable levels of separation from their healthy counterparts [1].

Cancer is a disease of the old age. The temporal patterns of ageing and cancer appear indeed to be perfectly superimposed: cancer is a rare occurrence in the young and becomes more and more common with age progression. The most common explanation of this phenomenon is based on the "multi-hit" hypothesis, according to which multiple "hits" to DNA are necessary to cause cancer. In this perspective, the chance that a cell accumulates the number of hits required for transformation increases with the age of the individual: thus, the late onset of cancer simply reflects the time necessary for a series of rare events to occur. The fact that a 6th power law fits well statistical data on cancer prevalence and age, seems to suggest that six independent hits are on average required for carcinogenesis.

5.3 Interpretation of carcinogenesis

As we showed in section 4, at the end of an individual's development (when reproduction occurs) many driver cells are present, which have not been activated: such driver cells rep-

resent a reservoir of change events that can potentially occur in the ageing period (i.e. after reproduction). Since the effects of these events on the survival chances of the individual's progeny are weak and steadily decreasing, they are subject to a correspondingly low evolutionary pressure. As a result, they tend to have a pseudo-random nature and their impact on the individual's health are more likely to be detrimental than beneficial. Overall, their cumulative effect translates to a slow, "benign" decrease of the individual's health, that we call ageing.

Now, the stage for a dangerous scenario is set if a fault arises in one of such ageing driver cells, affecting the mechanism used by the cell to generate new MOC values. Let us assume that in a driver cell (mother cell) the MOC-MOS match results in a proliferation event that creates new normal cells. Some of these cells are later turned into driver cells (daughter cells). The device responsible for MOC generation used by the daughters (the MOC generator) is inherited from their mother: if it is damaged, the damage will be present also in the daughters, affecting their capacity to generate new MOC values.

This device can be damaged in many ways: in one possible variant the damage can impede the creation of the new element which is responsible for differentiation (Fig. 15). Therefore one or more daughter driver cells will end up having the same MOC value as the mother (Fig. 16). Since these driver cells have the same MOC as the mother and obviously the same genome, the same MOC-MOS match that triggered a proliferation in the mother is bound to occur also in the daughters, giving rise to an identical proliferation. These proliferations triggered in the daughters are destined to produce granddaughter cells with the same fault in the MOC generator device. The result of this scenario is a chain of proliferation events, the mark of carcinogenesis. The engine of the process is a set of "cancer driver cells", which are continuosly and dinamically created, and all contain the initial damage to the MOC generator. The continuous creation of cells able to sustain the carcinogenic process seems to be confirmed by a recent study [14].

In section 2, we have hypothesised that the MOC generation device in biological systems corresponds to the pathways dedicated to the identification, transduction and processing of morphogens released in the external environment. The output of the MOC generator is a new transcription factor, which is added to the regulatory milieu of the cell. Therefore, a damage to the MOC generator corresponds in biology to mutations affecting genes involved in the pathways dedicated to the processing of morphogens (in short: morphogen-processing pathways and genes, or MP pathways and MP genes). Interestingly, these genes are frequently mutated or aberrantly activated in cancer [16, 24].

In ET, the set of MOC values necessary to induce differentiation in a given part of the structure (e.g. the orange cells in Fig. 6) is different from the set of MOC values needed to induce differentiation in a different part (e.g. the red cells in the same Figure). If we translate this to biology, the same must be true for tissues and MP pathways. Given, for instance, four MP pathways, (MP1, MP2, MP3, MP4), a subset of the pathways, say (MP1, MP2), could be necessary to generate the transcription factors needed for skin differentiation, and another subset, say (MP2, MP3, MP4), could be necessary to generate the transcription factors needed for gut differentiation. If a sufficient number of MP genes in a pathway are rendered non-functional (through mutations), the differentiation mechanism stops working. This situation is hypothesised to correspond to initiation, the first stage of carcinogenesis.

In ET the actual proliferation of the damaged driver cell, which corresponds to the biological stage of promotion, occurs once all conditions required for gene activation are met. In particular, the activation of the developmental gene is postponed until the clock reaches the timer value. It is worth noting that, if the cell does not proliferate, the effects of the damage to the MOC generator do not manifest themselves, even if present. This can lead to a long latency period, in which the potential for carcinogenic proliferation is present but not elicited.

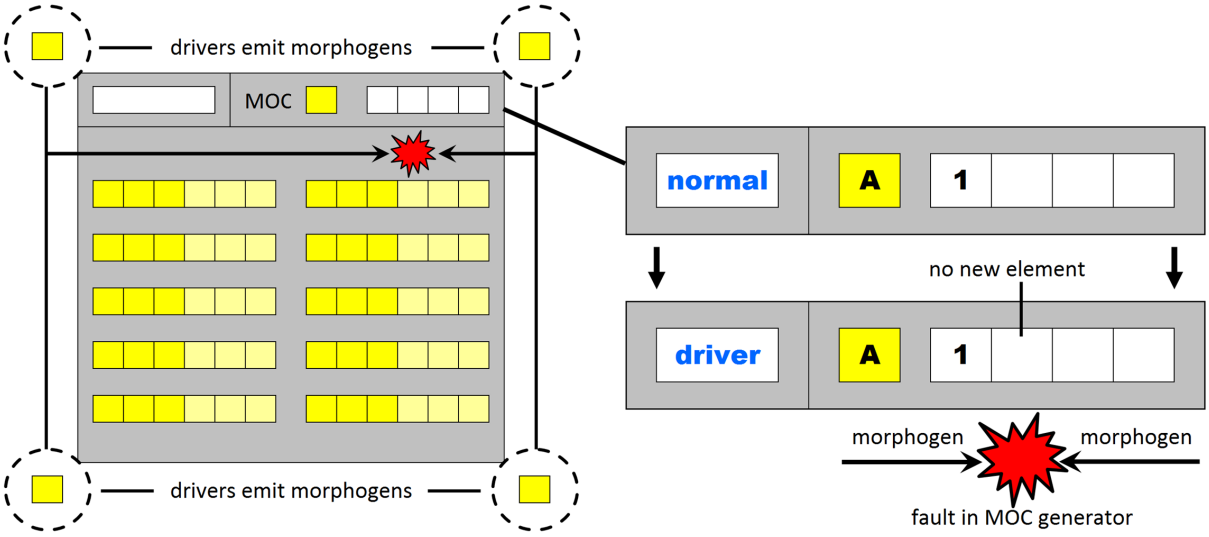

Figure 15: Carcinogenic generation of a new MOC value in a newly formed driver cell. A damage inside the MOC generator impairs the differentiation process. As a result, the MOC value of the daughter cell is the same as the mother.

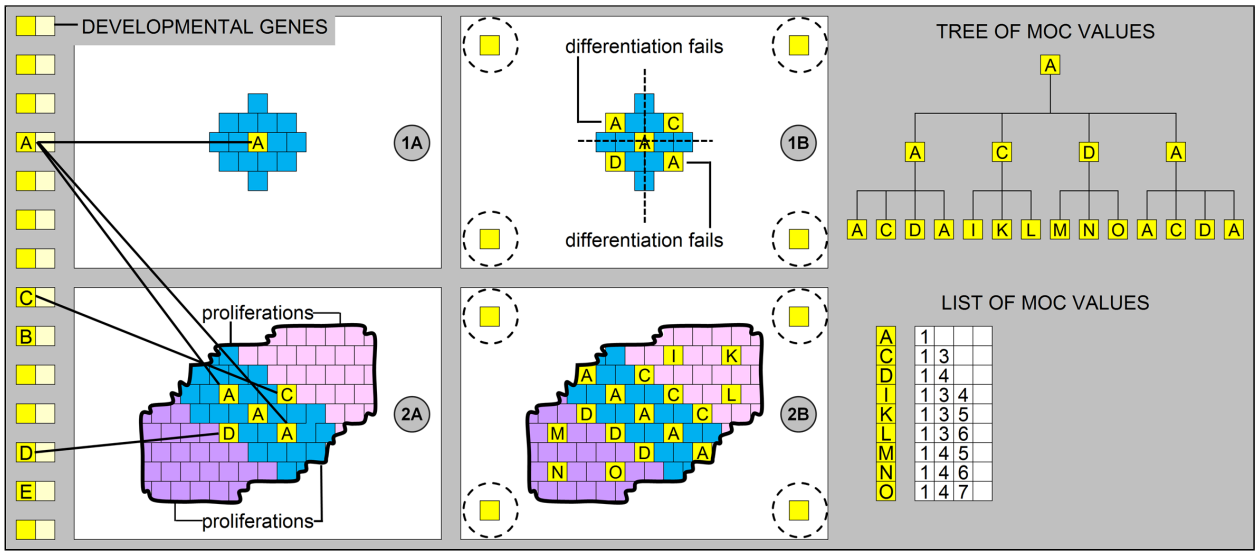

Figure 16: Generation of new driver cells during a carcinogenic proliferation event. The Figure reports the tumorigenic version of the development shown in Fig. 7. Thanks to a fault in the MOC generation device of the mother, one or more daughter driver cells have the same MOC value as the mother (value A in panel 1B). Since this MOC value triggers a proliferation for the mother (panel 1A), the same holds true for the daughters (panel 2A), for the granddaughters, etc., leading to an endless chain of proliferation events. As a result the population of blue (immature) cells tends to increase over time. Other cell types (pink and purple), which display higher differentiation and were generated during normal development, are still present.

Finally, progression, the third and last phase of carcinogenesis, is hypothesised to be driven by further mutations occurring to other genes, which confer additional powers to the already transformed cells, e.g. the capacity to infiltrate tissues and to produce distant metastases. We can now analyse how our explanation interprets the experimental evidence.

The proposed model is coherent with the theory of cancer stem cells, which foresees that cancers are sustained by a small subset of cells. In ET both development and cancer are caused by change events orchestrated by driver cells, which are fewer than normal cells by orders of magnitude. Driver cells, as we have seen, correspond to biological stem cells intended as cells capable of self-renewal and generation of cells destined to become specialised cells.

If we recall that the set of transcription factors necessary for differentiation is tissue-specific and assume that transcription factors are generated by MP genes, it is not surprising to find that the set of mutated MP genes is cancer-specific. A fully-working differentiation machinery requires all MP genes to be functioning. With reference to the example reported above (where MP1 and MP2 are necessary for skin differentiation), if, say, MP1 is damaged, the daughter driver cells become less different from their mother; if also MP2 is damaged, the transcription factor responsible for differentiation stops being produced and the daughter driver cells become equal to their mother. It is worth noting that, while full differentiation occurs in a unique way (all involved MP pathways have to be functional), a reduced form of differentiation can be realised in many different ways, corresponding to all possible combinations of non-functional MP pathways.

cancer type	bladder	brain	breast	colon
MP pathways involved	123	124	134	234
combinations of damaged pathways	1	1	1	2
	2	2	3	3
	3	4	4	4
	12	12	13	23
	13	14	14	24
	23	24	34	34
	123	124	134	234

The table above gives an overview of a hypothetical involvement of four MP pathways (1, 2, 3 , 4) in four organs / cancer types (bladder, brain, breast, colon). For each organ three MP pathways are hypothesised to be necessary for differentiation. The rows labelled “combinations of damaged pathways” list all possible combinations of damaged pathways which lead to failed differentiation, and hence tumour formation, in the relevant organ. In the case of breast cancer, for instance, pathways 1, 3 and 4 have to functional for full differentiation to occur. Different combinations of damaged pathways (1, 3, 4, 13, 14, 34, 134) cause impaired differentiation and may correspond to tumours of different grades. Considering that a damage in each MP pathway can be caused by mutations to a number of MP genes, this scheme can help to explain the heterogeneity of the mutational patterns observed.

The presence in tumours of cells having different degrees of differentiation and/or of different cell types is another fact which is easily accounted for in our model. As a matter of fact, along with driver cells bearing the same MOC value of the mother (the key feature of carcinogenesis), other driver cells with a “normal”, differentiated MOC value may be present. This because the combination of MP pathways necessary for the generation of their MOC values has not been affected by mutations. It is quite natural to think that the grade and the rate of growth of a tumour are linked to the share of driver cells having the same MOC value as the mother.

The proposed theory provides also a quite straightforward explanation for the relation between cancer prevalence and age. Indeed, the very same events triggered in the ageing period are hypothesised to contribute to the ageing phenomenon (if the MOC generation device is intact) or give rise to a tumour (if the MOC generation device is damaged). As a result, in our model the temporal patterns of ageing and cancer are coincident. The model sheds light also on the phenomenon of latency, i.e. the fact that the exposure to mutagenic substances (e.g. tobacco smoke) and the appearance of the related tumour (e.g. lung cancer) can be events separated by many years. As a matter of fact, even if the damage to the driver cell's differentiation device occurs early in life, for its effects to become manifest we need to wait until a proliferation event is triggered in the cell by a developmental gene: if the timer of the gene is set to 60 years of age, the tumour will not appear until that moment.

6 Conclusions

6.1 A new approach to the study of biology

In the so-called post-genomic era, it has become clear how the functioning of biological systems relies on hugely complicated networks of genes, proteins, chemical reactions. The most common approach to untertake the study of this complex matter can be described as "divide et impera" or "bottom-up". This translates to a narrowing of the field of expertise of many researchers, who have a very detailed knowledge in specific domains, and is also reflected in the multiplication of scientific journals specialised in specific sectors and sub-disciplines. Given the complexity of the subject, this approach has indeed been considered as the only possible one. Using a metaphor, we could say that biologists are trying to put together a huge puzzle, made up of hundreds of thousands of individual pieces, being each biologist or research laboratory concerned with a small part of the puzzle only.

Another consequence of the complexity of biological systems is the enormous flow of biological data produced by modern analysis techniques. In this context, the study of biology has become increasingly dominated by computer science. Examples of this trend are represented by the use of hardware and software tools for the sequencing of genomes of the most diverse species, the employment of specific software programmes for the automatic identification of gene sequences, the use of techniques of statistical analysis for the search of regular patterns in base sequences within genes, amino acid sequences within proteins, etc. The application of computer science tools to the analysis of biological data has given origin to a specific discipline called Bioinformatics.

One reason why the "bottom-up" approach has been considered the only viable one is that biological systems are believed to be inherently more complex than any other systems present in nature or man-made. Mechanical systems, which are effectively described by means of mathematical equations, are by comparison much simpler. On the other hand, we can imagine what the results would be is one tried to apply the same "puzzle" approach to the study of, say, statistical mechanics. In physical systems an astonishing number of sub-atomic particles is present and one such approach would mean to study the properties of individual particles in the hope to understand the functioning of the system at the macroscopic level. If such an approach would have been employed, the study of statistical mechanics would have faced difficulties similar to those presently encountered by modern biology.

The model presented in this work has been constructed with a very different approach, that can be defined "top-down": first, we have drawn the general architecture of the model at a high level, on the base of known biological elements; subsequently, we have added additional elements, not necessarily known, in order to produce interesting behaviours through computer simulations; finally, we have come back to biology, trying to guess which real biological elements play the role of the additional elements. As a consequence the model may (and does) contain elements not necessarily adherent to current knowledge, but which can become a suggestion for biologists to look into new, previously unexplored directions. Using the metaphor of the puzzle, this top-down approach corresponds to providing the picture of the puzzle, and use it as a guide to position the individual pieces.

6.2 Experiments to prove the theory

The most important feature of the model is the presence of driver cells. Therefore, it appears clear that the first step towards proving this theory is the identification of biological driver cells. According to ET, cells tend spontaneously to form a hierarchy, made up of normal cells and driver cells. Whenever a cellular proliferation takes place, new driver cells are created so that a uniform distribution of driver cells is always maintained in the cellular system.

The mechanism to create new driver cells is hypothesised to rely on the diffusion of chemical messengers from existing driver cells. This mechanism is known to play a role in the formation and maintenance of the "cancer driver cells" (i.e. cancer stem cells): efforts should be devoted to identifying the same mechanism also in healthy tissues during development and regeneration.

In a series of experiments conducted in the first decades of the 20th century, Hans Spemann demonstrated the potential of an area of the embryo, when transplanted into a second embryo, to induce the formation of a specific structure (the notochord) in an ectopic position. This area, called by Spemann "organiser", has ever since been referred to as "Spemann's organiser". The explanation of such experimental evidence by ET is straightforward: the transplanted area contains a driver cell which is the precursor of the notochord. The experiment of Spemann could be repeated, with the objective to identify such driver cell, keeping in mind that the generation of new driver cells is a dynamic process. The theory described foresees that natural driver cells are present in all organs of an organism's body, for the entire duration of the organism's life. Once natural driver cells have been correlated to specific biochemical markers, experiments should be aimed at finding such cells in the adult.

As we have seen, Germline Penetration implements a flow of genetic information from somatic cells to germline cells, to be passed on to future generations: as such, it can be considered the carrier of a transposon-mediated inheritance device. We mentioned how the biological phenomenon of sperm-mediated gene transfer could indeed represent the mechanism used by nature to implement the final stage of Germline Penetration, namely the phase in which new regulatory sequences enter the germline genome. The next logical step would be to try to verify the first part of the process, the one in which the TEs exit the driver cells in which they are hosted to reach the circulatory system and make their way towards the germline.

As we noted, while the experimental evidence reported in [23] suggests that the spreading of new TEs into the genome of a lineage and the occurrence of major changes in the lineage are simultaneous events, our model predicts that the spread of transposons is an event which immediately (in evolutionary terms) follows the evolutionary change. If the sensitivity of the analysis techniques used to estimate the age of DNA sequences were able to prove such prediction, this would represent a clear indication in favour of the proposed model.

6.3 Final remarks

Epigenetic Tracking is a model of systems of biological cells, able to generate arbitrary 3-dimensional cellular structures of any kind and complexity (in terms of number of cells, number of colours, etc.) starting from a single cell. If the complexity of such structures is interpreted as a metaphor for the complexity of biological structures, we can conclude that this model has the potential to generate the complexity typical of living beings. It has been shown how the model is able to reproduce a simplified version of some key biological phenomena such as development, the presence of "junk DNA", the phenomenon of ageing and the process of carcinogenesis. This model links the properties and behaviour of genes and cells to the properties and behaviour of the organism, describing and interpreting the said phenomena with a unified framework: for this reason, we think it can be proposed as a model for mutlicellular biology. Future work will be aimed at closing the gap with molecular biology, mapping the model variables to individual genes and chemical elements.

References

[1] D. Bonnet and J.E. Dick. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nature Medicine*, 1997.

[2] L. Collier and D. Largaespada. Transposable elements and the dynamic somatic genome. *Genome Biology*, 2007.

[3] E.M. De Robertis. Spemann's organizer and self-regulation in amphibian embryos. *Nature Reviews Molecular Cell Biology*, 2006.

[4] A. Fontana. Epigenetic tracking, a method to generate arbitrary shapes by using evo-devo techniques. In *Proceedings of EPIROB*, 2008.

[5] A. Fontana. Epigenetic tracking: biological implications. In *Proceedings of ECAL*, 2009.

[6] A. Fontana. An artificial life model for carcinogenesis. In *Proceedings of ALIFE 12*, 2010.

[7] A. Fontana. Devo co-evolution of shape and metabolism for an artificial organ. In *Proceedings of ALIFE 12*, 2010.

[8] A. Fontana. A hypothesis on the role of transposons. *Biosystems*, 2010.

[9] A. Fontana. *Epigenetic Tracking: an evo-devo approach to generate 3D simulated structures*. Phd in computer science, Technical University of Gdansk, 2012.

[10] A. Fontana and B. Wrobel. A model of evolution of development based on germline penetration of new "no-junk" dna. *Genes*, 2012.

[11] L.S. Frost, R. Leplae, A.O. Summers, and A. Toussaint. Mobile genetic elements: the agents of open source evolution. *Nature Reviews of Microbiology*, 2005.

[12] L.A. Gavrilov and N.S. Gavrilova. Evolutionary theories of aging and longevity. *The Scientific World Journal*, 2002.

[13] T.R. Gregory. Genome size and developmental complexity. *Genetica*, 2002.

[14] P.B. Gupta, C.M. Fillmore, G. Jiang, S.D. Shapira, K. Tao, C. Kuperwasser, and E.S. Lander. Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell*, 2011.

[15] B.L.M. Hogan. Morphogenesis. *Cell*, 1999.

[16] C. Karamboulas and L. Ailles. Developmental signaling pathways in cancer stem cells of solid tumors. *Biochimica et Biophysica Acta*, 2013.

[17] A. Knudson. Mutation and cancer: statistical study of retinoblastoma. *PNAS*, 1971.

[18] S. Kumar and P.J. Bentley. *On growth, form and computers*. Academic Press, 2003.

[19] A. Lindenmayer. Mathematical models for cellular interaction in development. *Journal of Theoretical Biology*, 1968.

[20] C.B. Lowe, G. Bejerano, and D. Haussler. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *PNAS*, 2007.

[21] B. McClintock. The origin and behavior of mutable loci in maize. *PNAS*, 1950.

[22] A.R. Muotri, M.C.N. Marchetto, N.G. Coufal, and F.H. Gage. The necessary junk: new functions for transposable elements. *Human Molecular Genetics*, 2007.

[23] K.R. Oliver and W.K. Greene. Transposable elements: powerful facilitators of evolution. *Bioessays*, 2010.

[24] T. Reya and H. Clevers. Wnt signalling in stem cells and cancer. *Nature*, 2005.

[25] T. Reya, S.J. Morrison, M.F. Clarke, and I.L. Weissman. Stem cells, cancer, and cancer stem cells. *Nature*, 2001.

[26] R. Ruddon. *Cancer Biology*. Oxford University Press, 2009.

[27] K. Smith and C. Spadafora. Sperm-mediated gene transfer: applications and implications. *Bioessays*, 2005.

[28] C. Stern, J. Charit, J. Deschamps, D. Duboule, A.J. Durston, M. Kmita, J.F. Nicolas, I. Palmeirim, J.C. Smith, and L. Wolpert. Head-tail patterning of the vertebrate embryo: one, two or many unresolved problems? *International Journal of Developmental Biology*, 2006.

[29] A. Turing. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society*, 1952.

[30] R.A. Van den Bussche, J.L. Longmire, and R.J. Baker. How bats achieve a small c-value: frequency of repetitive dna in macrotus. *Mammalian genome*, 1995.

[31] J.E. Visvader and G.J. Lindeman. Cancer stem cells in solid tumours: accumulating evidence and unresolved questions. *Nature Reviews Cancer*, 2008.

[32] C.H. Yeang, F. McCormick, and A. Levine. Combinatorial patterns of somatic gene mutations in cancer. *FASEB Journal*, 2008.